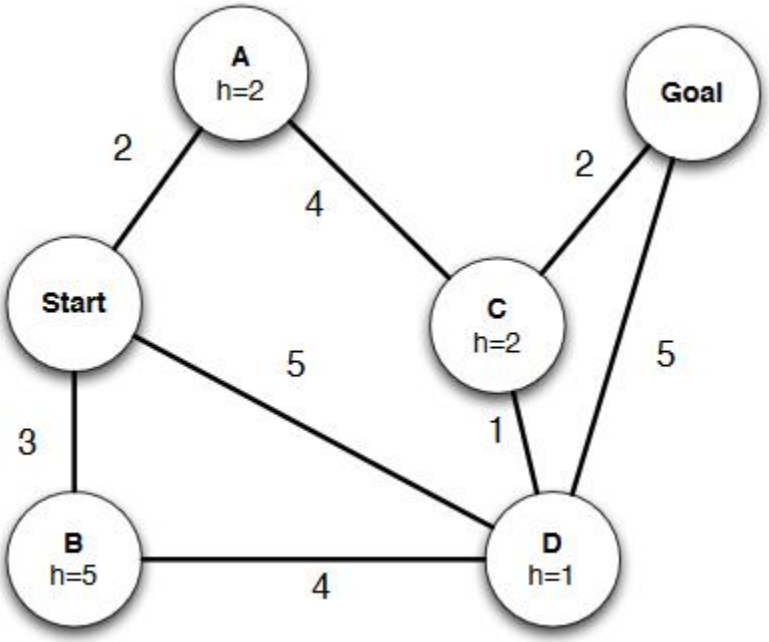
Course Title: **Artificial Intelligence and Expert Systems** Course Code: CCE3219 3rd year
Date: 9.4.2016 (Second term) Allowed time: 1 hrs

Answer Model:

Question No. 1 (10 marks- 2 marks for each point)

1. **For each of the following graph search strategies, work out the order in which states are expanded, as well as the path returned by graph search. In all cases, assume that states with earlier alphabetical order are expanded first. The start and goal state are S and G, respectively. Remember that in graph search, a state is expanded only once. (5 marks)**



a) Depth- first search.
States Expanded: Start, A, C, D, B, Goal
Path Returned: Start-A-C-D-Goal
b) Breadth- first search.
States Expanded: Start, A, B, D, C, Goal
Path Returned: Start-D-Goal
c) Uniform cost search.
States Expanded: Start, A, B, D, C, Goal
Path Returned: Start-A-C-Goal
d) Greedy search with the heuristic h shown on the graph.
States Expanded: Start, D, Goal
Path Returned: Start-D-Goal
e) A* search with the same heuristic.
States Expanded: Start, A, D, C, Goal
Path Returned: Start-A-C-Goal

2. Prove each of the following statements: (3 marks- 1 mark for each point)

a) Breadth-first search is a special case of uniform-cost search.

Breadth- first search is a special case of Uniform-cost search when all step costs are equal.

b) Breadth-first search, depth-first search, and uniform-cost search are special cases of best-first search.

Breadth-first search is best-first search with f(n) = depth(n).
Depth-first search is best-first search with f(n) = - depth(n).
Uniform-cost search is best-first search with f(n) = g(n).

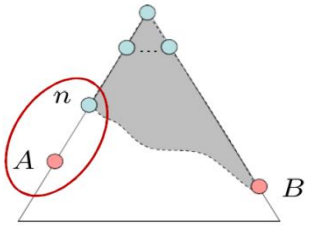
c) Uniform-cost search is a special case of A* search.

Uniform-cost search is A* search with h(n) = 0.

3. Prove that A* tree search with admissible heuristic is optimal. (2 marks)

The Proof

Assume B is on the fringe and some ancestor *n* of A is on the fringe, too (may be A!). Assume *n* will be expanded before B



$f(n) = g(n) + h(n)$ Definition of f-cost
$f(n) \leq g(A)$ Admissibility of h
$g(A) = f(A)$ h = 0 at a goal

Then f(n) <= f(A) (1)

$g(A) < g(B)$ B is suboptimal
$f(A) < f(B)$ h = 0 at a goal

Then f(A) < f(B) (2)

from (1) and (2) Then $f(n) \leq f(A) < f(B)$

n expands before B. All ancestors of A expand before B. A expands before B.

Then A*search is optimal

Question No. 2 **(10 marks)**

For each of the following, please circle the letter introducing the best answer. (Check all that apply.) Explain your answer. Each one worth one degree

1. Suppose you are working on stock market prediction, and you would like to predict whether or not a particular stock's price will be higher tomorrow than it is today. You want to use a learning algorithm. Which one of the following algorithms is appropriate?

a) Regression
b) Classification
c) Clustering
d) Reinforcement learning

The correct choice b

Classification is appropriate when we are trying to predict one of a small number of discrete-valued outputs. Here, there are two possible outcomes: That the stock price goes up (which we might designate as class 0, say) or that it does not (class 1).

2. You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

a) Treat both as classification problems.
b) Treat problem 1 as a classification problem, and problem 2 as a regression problem.
c) Treat problem 2 as a classification problem, and problem 1 as a regression problem.
d) Treat both as regression problems.

The correct choice c

Regression is appropriate when we are trying to predict a continuous-valued output, such as in problem 1, the items that will be sold over the next 3 months.

Classification is appropriate when we are trying to predict one of a small number of discrete-valued outputs. In problem 2, there are two possible outcomes: hacked/compromised

3. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is E?

a) The process of the algorithm examining a large amount of historical weather data.
b) The weather prediction task.
c) The probability of it correctly predicting a future date's weather.
d) None of these.

The correct choice a

Examining a large amount of historical weather data considers as learning from experience

4. Suppose you have a dataset with m=50 examples and n=100000 features for each example. You want to use multivariate linear regression to fit the parameters to our data. Should you prefer gradient descent or the normal equation?

a) The normal equation, since gradient descent might be unable to find the optimalθ.
b) The normal equation, since it provides an efficient way to directly find the solution.
c) Gradient descent, since it will always converge to the optimal θ.
d) Gradient descent, since $(X^TX)^{-1}$ will be very slow to compute in the normal equation.

The correct choice d

With n=200000 features, you will have to invert a 200000x200000 matrix to compute the normal equation. This is a complex inversion, so the gradient decent is efficient.

5. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

a) The cluster assignment step, where the parameters $C^{(i)}$ are updated.
b) Move the cluster centroids, where the centroids μ_k are updated.
c) Feature scaling, to ensure each feature is on a comparable scale to the others.
d) Using the elbow method to choose K.

The correct choice a and b

a.....The assignment step is the first step of the K-means loop.
b.....The cluster update is the second step of the K-means loop.

6. Suppose you have an un-labelled dataset. You run K-means with 50 different random initializations, and obtain 50 different clusters of the data. What is the recommended way for choosing which one of these 50 clusters to use?

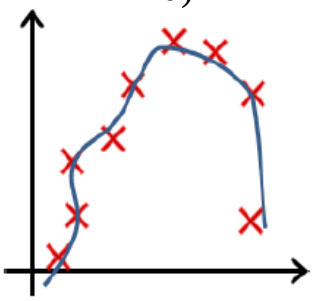
a) The answer is ambiguous, and there is no good way for choosing.
b) Compute the distortion function and pick the one that minimizes it.
c) Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.
d) The only way to do so is if we also have labels for our data.

The correct choice b

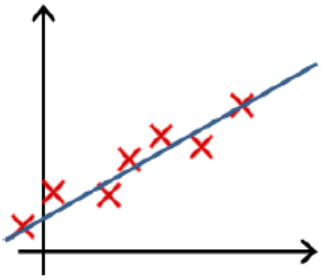
A lower value for the distortion function implies a better clustering, so you should choose the clustering with the smallest value for the distortion function.

7. In which one of the following figures do you think the hypothesis has overfit the training set?

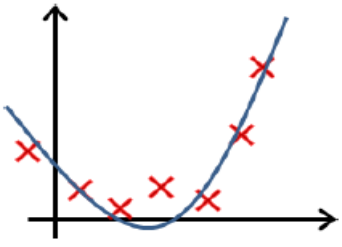
a)



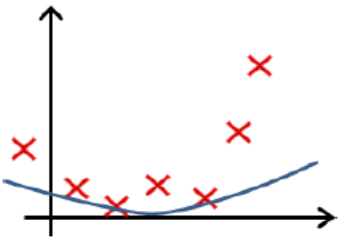
b)



c)



d)



The correct choice a

If we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples (predict prices on new examples).

8. For which of the following tasks might K-means clustering be a suitable algorithm?

a) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
b) Given many emails, you want to determine if they are Spam or Non-Spam emails.
c) Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.
d) Given a database of information about your users, automatically group them into different market segments.

The correct choice c and d

Examples of un-labeled data sets

9. Suppose you ran logistic regression twice, once with λ=1 and once with λ =1. One of the times, you got parameters $\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$, and the other time you got $\theta = \begin{bmatrix} 1.37 \\ 0.51 \end{bmatrix}$. However, you forgot which value of λ corresponds to which value of θ. Which one do you think corresponds to λ=1?

a) $\theta = \begin{bmatrix} 1.37 \\ 0.51 \end{bmatrix}$

b) $\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$

The correct choice a

When λ is set to 1, we use regularization to penalize large values of θ. Thus, the parameters θ obtained will in general have smaller values.

10. A navigation system that first considers all possible routes to the destination, and then selects the shortest route is described as:

a) Reflex agent.
b) Planning agent.
c) Multi-agent
d) Substituted Agent

The correct choice b

Agent that plans ahead. It asks "what if", takes decisions based on (hypothesized) consequences of actions. It must have a model of how the world evolves in response to actions. It must formulate a goal (test). It considers how the world WOULD BE.

Best wishes

Dr. Sherin El Gokhy